

FTSite: High accuracy detection of ligand binding sites on unbound protein structures

Chi-Ho Ngan, David Hall, Brandon Zerbe, Laurie E. Grove, Dima Kozakov, Sandor Vajda

Supplementary figures and text:

Supplementary Methods	
Supplementary Results	
Supplementary Discussion	
Supplementary Figure 1	Small molecules used as probes by FTSite to identify ligand binding sites.
Supplementary Figure 2	Top ranked predictions of binding sites identified by FTSite in the unbound structures of QSiteFinder test set
Supplementary Figure 3	The top ranked FTSite prediction of the ligand binding site, has excellent coverage of the ligand.
Supplementary Figure 4	Cases in which the ligands are large and/or are branched.
Supplementary Figure 5	Cases in which FTSite succeeded in finding the ligand binding sites but other methods had challenges.
Supplementary Figure 6	Cases in which FTSite was unable to identify the ligand binding site using only the top ranked prediction.
Supplementary Figure 7	User interface for the FTSite server.
Supplementary Figure 8	Online interface for the user to view the results after the analysis is done.
Supplementary Table 1	Contact energies for atoms from the protein and atoms from the 16 small molecule probes used by FTSite.
Supplementary Table 2	Comparing the performances of FTSite to those of QSiteFinder and the method of Morita et al. on the QSiteFinder test set.
Supplementary Table 3	Prediction of ligand binding site for the proteins in the LIGSITE ^{CSC} test set.
Supplementary References	

Supplementary Methods

Outline of the FTSite algorithm

Step 1: Grid-based sampling of the protein surface using FFT. Protein structures are downloaded from the Protein Data Bank¹. All bound ligands including water molecules are removed prior to the calculations. FTSite uses 16 small molecule probe types (Supplementary Fig. 1) to sample the protein surface. Grid-based Fast Fourier Transform (FFT) samples the protein exhaustively with 10^9 docked probe positions on the protein surface; this is discussed in greater detail afterwards. This algorithm requires only the atomic coordinates of the probe and those of the protein structure; no *a priori* information about the ligand binding site is required. The best 2,000 poses with the lowest energies for each probe type are retained.

Step 2: Post-FFT clustering to discard spurious probe clusters. For each probe type, the 2,000 retained poses are clustered using a simple greedy algorithm. Based on biophysical arguments we select the lowest energy pose as the center of the first cluster, and add all poses within 4 Å center-to-center distance from it as cluster members. All clustered poses are removed, and we repeat the same steps to form the second and the subsequent clusters until all poses are clustered. Clusters with less than 10 probes are removed, and the 6 largest clusters are retained for further analysis.

Step 3: Off-grid minimization and re-scoring. The energy of each retained protein-probe complex is minimized using the CHARMM² potential with the Analytic Continuum Electrostatic (ACE)³ model representing the electrostatics and solvation terms as implemented in version 27 of CHARMM². The algorithm uses the polar-hydrogen-only parameter set from version 19 of CHARMM². The energy minimization is performed using a limited memory Broyden–Fletcher–Goldfarb–Shannon (L–BFGS) method in which heavy atoms of the protein are held fixed, while the polar hydrogen atoms of the protein and all atoms of the probes are free to move. Poses with positive energies after minimization are discarded.

Step 4: Generating consensus clusters. Following the energy minimization we re-cluster the resulting probe poses. As in step 2 we select the lowest energy pose as the center of the first cluster, but use 4 Å full-atom pairwise RMSD as the clustering radius. After all probes are clustered and clusters with less

than 10 members are discarded, the clusters are ranked on the basis of the Boltzmann averaged energy, and the 6 lowest energy clusters are retained for every probe type. Consensus clusters are generated by grouping probe clusters with clusters centers within 4 Å. The centers of the resulting consensus clusters are fixed, and the probe clusters are re-distributed such that each cluster center is closer to the center of its own consensus cluster than to the center of any other consensus cluster. Consensus clusters that overlap with an integral element of the intact protein such as a co-factor are discarded. A consensus cluster is considered to overlap with a co-factor if their volume overlap exceeds 80% of the consensus cluster.

Step 5: Ranking consensus clusters. The algorithm ranks the consensus clusters by the number of non-bonded contacts between the protein and all probes of the consensus cluster. A residue of the protein and a probe are considered to be in contact if any atom of the residue is less than 4 Å from any atom of the probe. A residue is considered to be in contact with a consensus cluster if it is in contact with any of its probes. After selecting the contact residues for a consensus cluster we re-evaluate the number of contacts by adding also interactions with probes that are within 4 Å but are not part of the original consensus cluster. The resulting numbers are normalized using the overall number of contacts for all probes, and used for ranking the consensus clusters.

Step 6: Identification of putative ligand binding sites. To identify the putative ligand binding site the algorithm first selects the consensus cluster with the highest number of contacts. This cluster is then expanded by adding any neighboring consensus cluster if the center of any of its probe is closer than 3.5 Å to the center of any probe in the consensus cluster. The protein residues that are within 4 Å of the expanded consensus cluster constitute the top prediction of the binding site. The first consensus cluster is then removed, and the procedure is repeated using the next consensus cluster with the highest number of contacts to identify lower ranked predictions of the ligand binding site.

The Fast Fourier Transform (FFT) correlation approach to mapping

In Step 1 we perform exhaustive evaluation of an energy function in the discretized 6D space of mutual orientations of the protein (receptor) and a small molecule probe (ligand). The center of mass of the receptor is fixed at the origin of the coordinate system. The translational space is represented as a grid of 0.8 Å displacements of the ligand center of mass, and the rotational space is sampled using 500 rotations

based on a deterministic layered Sukharev grid sequence which quasi-uniformly covers the space⁶.

The energy function describing the receptor-ligand interactions is defined on the grid and is expressed as the sum of P correlation functions for all possible translations α, β, γ of the ligand at a given rotation:

$$E(\alpha, \beta, \gamma) = \sum_p \sum_{i,j,k} R_p(i, j, k) L_p(i + \alpha, j + \beta, k + \gamma)$$

where $R_p(i, j, k)$ and $L_p(i, j, k)$ are the components of the correlation function defined on the receptor and the ligand, respectively. This expression can be efficiently calculated using P forward and one inverse Fast Fourier transforms, denoted by FT and IFT , respectively:

$$\begin{aligned} E(\alpha, \beta, \gamma) &= IFT\left(\sum_p \{FT^*\{R_p\}FT\{L_p\}\}\right)(\alpha, \beta, \gamma) \\ FT\{F\}(l, m, n) &= \sum_{i,j,k} F(i, j, k) \exp^{-2\pi \mathbf{i}(li/N_1 + mj/N_2 + nk/N_3)} \\ IFT\{f\}(i, j, k) &= C \sum_{l,m,n} f(l, m, n) \exp^{2\pi \mathbf{i}(li/N_1 + mj/N_2 + nk/N_3)} \end{aligned}$$

where $\mathbf{i} = \sqrt{-1}$, N_1 , N_2 , and N_3 are the dimensions of the grid along the three coordinates, and $C = 1/(N_1 N_2 N_3)$. If $N_1 = N_2 = N_3 = N$, the efficiency of this approach is $O(N^3 \log(N^3))$ as compared to $O(N^6)$ when all evaluations are performed directly.

For each rotation of the ligand we generate the $FT(L_p)$ function on the grid and then calculate the sum of the correlation functions using the formula above, resulting in scoring function values for all possible translations. Since the function may have multiple minima, we retain the four lowest energy regions of the translational space for each rotation. To derive the first region we select the lowest energy solution, remove the surrounding volume of the 27\AA^3 cube, and repeat this step three more times. Finally, results from different rotations are collected and sorted.

Energy function in the FFT based grid docking step

The energy expression in Step 1 includes the simplified van der Waals energy E_{vdw} with attractive (E_{attr}) and repulsive (E_{rep}) contributions, the electrostatic interaction energy E_{elec} , an enclosure term E_{encl} describing the contributions from hydrophobic enclosures, and the statistical knowledge-based pairwise potential E_{pair} representing other solvation effects:

$$\begin{aligned}
 E &= E_{vdw} + w_2 E_{elec} + w_3 E_{encl} + w_4 E_{pair} \\
 E_{vdw} &= E_{attr} + w_1 E_{rep} \\
 E_{elec} &= \sum_{i=1}^{N_l} q_i \phi_{rPB} \\
 E_{pair} &= \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \varepsilon_{ij}
 \end{aligned}$$

where N_R and N_L denote the numbers of atoms in the receptor and the ligand, respectively. The coefficients $w_1=11.1$, $w_2=44.4$, $w_3=0.88$, and $w_4=3.33$ weight the different contributions to the scoring function based on calorimetric considerations and in agreement with the parameters used in the FTMap algorithm⁷.

Van der Waals energy. We use stepwise functions to represent the attractive and repulsive steric terms. The repulsive interactions are cut off at the van der Waals radius r_{vdw} plus 1.8 Å because we want the penalty function to be tolerant enough and to allow for differences between bound and unbound structures. The attractive part is truncated at 6 Å. On the grid, the functions describing the receptor and the ligand can be represented as follows.

$$\begin{aligned}
 R_p(l, m, n) &= -c_{l,m,n} + w_1 r_{l,m,n} \\
 L_p(l, m, n) &= \begin{cases} 1 & \text{if } (l, m, n) \ni (a_j \in J) \\ 0 & \text{otherwise} \end{cases}
 \end{aligned}$$

where $c_{l,m,n}$ is the number of atoms that are at the distance $d < r < D$ from the grid point (l, m, n) , $r_{l,m,n}$ is the number of atoms that are at the distance $r < d$ from the same grid point, and $(l, m, n) \ni (a_j \in J)$ means that the grid point (l, m, n) overlaps with atom a_j of atom type J . As mentioned, $D = 6$ Å and $d = r_{vdw} + 1.8$ Å. Thus, the correlation of R_p and L_p provides a shape complementarity term representing both repulsive and attractive interactions, the former for the distances $r < d$, and the latter

in the range $d < r < D$.

Electrostatic interactions. We approximate the electrostatic energy as the interaction energy between the electrostatic potential ϕ_{rPB} of the solvated protein and the atomic charges q_i of the probe. Thus, the influence of the probe on the electrostatic potential of the protein-solvent system is neglected, assuming that the probe is small and not strongly charged. Using a dielectric continuum model with low ion concentration (corresponding to 0.1 mol salt concentration), the electrostatic potential of the solvated protein is calculated by solving the linearized Poisson-Boltzmann equation

$$\nabla[\epsilon(\vec{r})\nabla\phi_{rPB}(\vec{r})] - \kappa^2(\vec{r})\phi_{rPB}(\vec{r}) = -4\pi\rho(\vec{r}),$$

where $\epsilon(\vec{r})$, $\kappa(\vec{r})$, and $\rho(\vec{r})$ are the dielectric constant, the modified Debye-Hückel screening factor, and the fixed charge density of the protein, respectively. The dielectric boundary between the low dielectric protein region and the external bulk solvent is placed to account for the reduced water mobility and hence reduced polarization in binding site cavities. This is achieved by dividing atoms of the protein into "cavity" and "non-cavity" sets. Atoms are considered "cavity" if they are not accessible to a large spherical probe of 5.75 Å radius. The size of the probe is selected to represent the typical dimensions of protein active sites. Each cavity atom is assigned a dielectric radius equal to its van der Waals radius plus 1.4 Å. In contrast, each non-cavity atom has a small fixed dielectric radius of 0.1 Å. These radii define a continuous surface that separates the low dielectric protein and its extension into the cavities ($\epsilon = 4.0$) from the bulk solvent ($\epsilon = 80.0$). We use the Poisson-Boltzmann module PBEQ (Beglov and Roux, unpublished) of CHARMM² to calculate the potential ϕ_{rPB} . The electrostatic energy is then expressed as the vector product of the functions

$$\begin{aligned} R_p(l, m, n) &= \phi_{rPB}(l, m, n) \\ L_p(l, m, n) &= \begin{cases} q_j & \text{if } (l, m, n) \ni (a_j \in J) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

defined on the receptor and on the ligand grids, respectively. The potential is truncated at 15.0 kcal/mol.

Pairwise statistical potential. The general form of a pairwise contact potential is

$$E_{pair} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \varepsilon_{ij}$$

For a pair of atoms a_i and a_j of types I and J , respectively, $\varepsilon_{ij} = \varepsilon_{IJ}$, where ε_{IJ} is the contact energy between atoms of types I and J , if $d < r_{ij} < D$; otherwise $\varepsilon_{ij} = 0$. We use the DARS (Decoys As the Reference State) potential that originally has been developed for protein-protein docking⁵, and has been extended to describe the interactions between proteins and the molecular probes considered here. The DARS parameters used in this work are listed below. In order to evaluate the energy function using Fast Fourier Transforms, it must be written as a sum of correlation functions. Based on the eigenvalue-eigenvector decomposition of the matrix of pairwise interaction coefficients ε_{IJ} , these coefficients can be written as

$$\varepsilon_{IJ} = \sum_{p=1}^K \lambda_p u_{pI} u_{pJ}$$

where λ_p is the p th eigenvalue of the interaction matrix, and u_{pI} is the I th component of the p th eigenvector. Each term in the eigenvalue - eigenvector decomposition represents an energy contribution proportional to the absolute value of the eigenvalue λ_p , and such contributions are independent due to the orthogonality of the eigenvectors. We have shown that restricting consideration to the first four terms yields around 10% error in the energy values, comparable to the error of representing the energies on a grid⁵. The energy term with the p th eigenvalue of the pairwise potential is defined by the correlation of the functions

$$\begin{aligned} R_p(l, m, n) &= \sum_{i=1}^{N_r} u_{pI} \delta_i \\ L_p(l, m, n) &= \begin{cases} u_{pJ} & \text{if } (l, m, n) \ni (a_j \in J) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where δ_i is 1 if atom i of the receptor is closer than 6 Å to the grid point (l, m, n) .

Enclosure term. Nonpolar enclosures disrupt water structure and create a favorable environment for ligand binding⁸. To represent this effect we place a Gaussian ball, with $\sigma = 10$ Å, at each grid point, and calculate its correlation with the C_α atoms of nonpolar residues. For each point in space, this function measures the fraction of the ball occupied by the nonpolar regions of the protein:

$$\begin{aligned} R_p(l, m, n) &= \sum_{i=1}^{N_R} \frac{1}{\sigma \sqrt{2\pi}} \exp\left(\frac{-r_{i,(lmn)}^2}{\sigma^2}\right) \\ L_p(l, m, n) &= \begin{cases} 1 & \text{if } (l, m, n) \ni A_L \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

where N_R is the number of atoms in the receptor, $r_{i,(lmn)}$ is the distance from atom i to grid point (l,m,n) , and A_L is an atom of the ligand probe. The correlation of R_p and L_p is large in nonpolar cavities, and it is small on protrusions and flat surfaces. This way the function, added to the energy expression with a negative sign, improves the sampling of the cavity regions.

Selection of the probe library

For all mapping calculations in this paper we used the 16 probe molecules shown in Supplementary Figure 1, which were inspired by original soaking experiments. These probes were selected to provide some diversity in shape and polarity. Most probes consist of a hydrophobic moiety and one or two polar groups (amine, amide, alcohol, ketone, urea, and ester), but we also included hydrophobic and aromatic compounds. Each group is represented with one or more simple members. Alcohols are particularly informative, since the OH group can act both as hydrogen bond donor and acceptor. All the probe parameters apart from DARS potential which is discussed later, including radii and charges were derived using version 19 of the CHARMM² potential.

DARS (Decoys As the Reference State) parameters for mapping

The DARS (Decoys As the Reference State) pairwise interaction potential was originally developed for protein-protein docking⁴. The novelty of the DARS approach is that we generate a large decoy set of docked conformations to be used as the reference state. To create the potential, we compare the frequency of contacts between two specific atom types in the native state to the frequency of contacts in the decoys. Supplementary Table 1 lists the DARS contact energies between each of the 18 atom types for proteins and each of the 67 heavy atom types of the 16 probes.

Test sets and success criteria

FTSite was evaluated on unbound structures from two test sets of proteins. The QSiteFinder set⁹ consists of both unbound and bound structures for 35 proteins. The LIGSITE^{CSC} set¹⁰ includes the same structures, plus unbound and bound pairs for 13 additional proteins that have been selected for assessing the binding site identification method called PASS¹¹. A number of methods have been tested using the two sets, for each set with somewhat different success criteria. Here we adopt the same criteria, since this will allow for comparing the performance of FTSite to that of other methods. For the QSiteFinder set⁹, the binding site is considered correctly predicted if at least 25% of the probes identifying the putative ligand binding site are within 1.6 Å of any ligand atom. This ensures that the site is reasonably sized, as substantially expanding the predicted site one could move too far from the ligand. For the LIGSITE^{CSC} set¹⁰ the criterion requires the geometric center of the predicted ligand binding site to be within 4.0 Å of any ligand atom. In addition to these criteria, this study introduces two additional quality measures, site coverage (SC) and ligand coverage (LC), defined as $SC=(VL \cap VS)/VS$ and $LC=(VL \cap VS)/VL$, respectively, where VL and VS denote the volume of the ligand and that of the predicted site, and \cap indicates the intersection of two volumes. The quality measure SC is in fact analogous to the success criteria used for the QSiteFinder set⁹, which thus requires an SC value exceeding 25%.

Supplementary results

Supplementary Table 2 compares the performance of FTSite, QSiteFinder⁹, and the method of Morita et al.¹² by providing site coverage for all bound/unbound protein pairs from the QSiteFinder set⁹ (successful predictions only). Supplementary Figure 2 shows top ranked predictions of binding sites identified by FTSite in the unbound structures of the QSiteFinder set⁹. The SC values for QSiteFinder⁹ and for the method of Morita et al.¹² are from the publication by Morita et al.¹² Supplementary Table 3 shows the distances from the geometric center of each putative ligand-binding site to the closest atom of the ligand (successful predictions only) for all bound/unbound protein pairs of the LIGSITE^{CSC} set¹⁰, as well as the site coverage (SC) and ligand coverage (LC) values.

There are obvious limitations to each of the criteria used by QSiteFinder⁹ and LIGSITE^{CSC} ¹⁰. In particular the bound ligands in the test sets may not be representative of all ligands because it is derived only from one bound structure. If the selected ligand is small, there can only be limited intersection of the putative ligand binding site with the ligand, which can result in a low SC value in spite of the correctly identified ligand binding site. On the other hand, if the ligand is much larger in volume than the predicted ligand binding site, then requiring an SC value of at least 25% may be too lenient. In particular, the QSiteFinder⁹ criterion would be relatively easy to satisfy for a small binding site accommodating only a fraction of a large ligand, although the prediction could be very inaccurate. Thus, adding ligand coverage (LC) as the second quality measure is very important if the predicted site overlaps only with a fraction of the ligand. On the other hand, the LIGSITE^{CSC} ¹⁰ criterion may be more restrictive but in some cases unphysical. If the ligand binding site is elongated and the ligand lies at one of its end, then the centers of geometry of the ligand and of the ligand binding site would necessarily be far apart, resulting in an apparently unsuccessful prediction.

Supplementary Discussion

Binding site identification methods used for comparison

A large number of ligand binding site identification methods have been published, which can generally be classified into geometry based and energy based methods, and may also involve the use of evolutionary information⁹⁻²⁵. The performance of FTSite has been compared to the performance of methods that have been evaluated either on the QSiteFinder⁹ or the LIGSITE^{CSC 10} test sets, in each case resulting in a ranked list of predicted binding sites. Some of these methods are geometry based including SURFNET¹³, POCKET¹⁴, LIGSITE¹⁵, LIGSITE^{CS 10}, CAST¹⁶, PASS¹¹, FPocket²⁰, PocketPicker²², DoGSite²³, and VICE²⁴. The energy based methods used for comparison include QSiteFinder⁹ and a similar method published by Morita et al¹². The STP method²⁵ uses residue triplet propensities extracted from protein structures, and since the approach is reminiscent of extracting a structure-based statistical potential, it can also be considered an energy based method. Finally, LIGSITE^{CSC 10} is an improved implementation of LIGSITE^{CS 10} where the authors used evolutionary information to improve their success rates in identifying ligand binding sites¹⁰. Therefore LIGSITE^{CSC 10} represents a combination of geometry based and evolutionary methods.

It is instructive to discuss some of the inherent limitations of the above methods, as the discussion may explain some sources of the improved predictive power by FTSite. A number of the methods are grid-based and are sensitive to grid-spacing or resolution, and in particular to the orientation of the protein in relation to the grid axes. Although the first step of FTSite is also grid-based, the method includes an off-grid minimization step that eliminates any noticeable dependence on grid placement. In some of the geometry-based methods it is challenging to accurately delineate free space from the ligand binding site. A number of methods use probe spheres to flood the protein surface and the clustering of the spheres to identify protein cavities and putative ligand binding sites. These methods typically struggle with identifying wide cavities unless spheres of greater radii are used. The natural drawback to the use of greater spheres is the “overflow” of spheres into neighboring cavities resulting in co-joined putative ligand binding sites that are imprecise. A number of methods require the computation of alpha shapes and the use of “discrete-flow” method to join neighboring cavities. These methods require the opening of a cavity to be smaller in circumference than any cross-section of the space, which holds for some ligand binding sites but not for others. An additional drawback is their tendency to join cavities buried inside the protein, thereby identifying continuous channels instead of a well-defined ligand binding site.

Several methods use training sets of proteins to build a classifier (shape descriptors, volumes etc.) for distinguishing ligand binding sites from unremarkable protein cavities. These classifiers tend to be

retrospective in nature, and hold limited information on the biophysical basis of protein-ligand interactions. These methods encounter significant challenges when applied on unbound proteins where the ligand binding sites are less well formed than in the bound forms of the proteins, or if there are significant conformational changes upon ligand binding. LIGSITE^{CSC}¹⁰ uses evolutionary based information on the re-ranking of putative ligand binding sites. The successful implementation of this method is contingent upon the quality of multiple sequence alignment tools, and the availability of sequences. FTSite circumvents many of these challenges by the direct biophysical modeling of protein-ligand interactions using molecular mechanics force fields. The consideration of protein geometry and physicochemical properties of a putative ligand binding site is implicit in the modeling.

Discussion of specific cases

The following is a selection of cases demonstrating the strengths and the occasional shortcomings of FTSite in the context of the two test sets used. First, cases in which FTSite has performed well and identified the entire ligand binding site correctly are discussed, followed by cases in which two of the top ranked predictions were needed to cover the entire site. Quality is assessed in terms of site coverage (SC) and ligand coverage (LC) values. Next we describe cases in which other energy based methods had difficulties (i.e., the ligand binding site was not among the 3 top ranked predictions), but FTSite worked well. Finally we discuss the few cases in which FTSite was unable to identify the correct ligand binding site as the top ranked prediction.

High-quality identification of binding sites. Supplementary Figure 3 shows four cases in which the top ranked prediction of the binding site agrees well with the ligand position in the bound form of the protein. In all these cases the ligand occupies almost the entire space predicted as the binding site, resulting in site coverage (SC) and ligand coverage (LC) values well above 90%. It is clear that in these proteins the binding site is well formed even in the unbound structure, and there are only mild conformational changes upon ligand binding. Therefore FTSite had no difficulty in locating the sites with high accuracy.

Partial identification of binding sites. Supplementary Figure 4 shows four cases in which the top ranked prediction overlaps only with some part of the ligand. These results are unsurprising given that the ligands span a large volume (Supplementary Fig. 4A and 4B), or extend in tangential directions out of the pocket (Supplementary Fig. 4C). In these cases, consensus clusters are located on the two distal ends of the ligands, and are not joined into a single consensus cluster. As the result, we need the two top ranked predictions of the binding site to trace out the ligand in its entirety. These cases are interesting as they demonstrate the concept in which a ligand binding site is identified by a collection of

consensus clusters, representing distinct hot spots. The amino acid residues in each hot spot contribute disproportionately to the binding free energy. Figure 4D demonstrates extension of the distal end of the ligand into the second putative ligand binding site; this hints at the feasibility of extending the ligand to interact with the second site, providing potentially important information for drug design.

FTSite outperforms other methods. There are a number of cases in which FTSite was able to identify the ligand binding site using the top ranked prediction, but other methods had challenges. In the cases selected (Supplementary Fig. 5A, 5B, and 5C) QSiteFinder⁹ was unable to correctly identify the ligand binding site using the 3 top ranked predictions. QSiteFinder⁹ found the putative ligand binding site for beta-amylase (Supplementary Fig. 5A) using the 4th predicted site, and for HIV-2 protease (Supplementary Fig. 5B) using the 7th predicted site. In the case of acetylcholinesterase (Supplementary Fig. 5C), the predicted ligand binding site has lower than 25% volume overlap with the ligand, and hence is not considered a successful prediction. In the case of beta-amylase, the loop formed by the residues V99, G100, and D101 significantly changes conformation, and closes down on the ligand in the bound form of the protein, resulting in a better defined ligand binding site. Similarly, in the unbound form of HIV-2 protease the ligand binding site is very open, and becomes well formed only upon ligand binding. Under these circumstances the interactions between the protein and the methyl probes used by QSiteFinder⁹ are diminished, which results in a lack of probe clusters and poor ligand binding site prediction. In contrast, FTSite samples the protein surface using small molecular probes that vary in size, shape, and polarity. Such probes extensively interact with the protein surface even when the hydrophobic spheres used by QSiteFinder do not, providing improved robustness of FTSite to conformational changes. The method developed by Morita et al.¹², which is very similar to QSiteFinder⁹, was unable to identify the ligand binding sites for beta-amylase (Supplementary Fig. 5A) and HIV-2 protease (Supplementary Fig. 5B) using any of its predictions. For trypsinogen (Supplementary Fig. 5D), Morita et al.¹² found the correct ligand binding site using the 6th predicted site. In all cases discussed here, the unbound and bound forms of the protein substantially differ in the region of the binding site, which makes finding the site based on the unbound structure more difficult.

FTSite negative results. In the two test sets, FTSite was unable to correctly identify the ligand binding sites using the top ranked prediction only for three proteins. These are 1A6U (unbound antibody Fv fragment), 1GCG (unbound glucose/galactose receptor), and 1ULA (unbound purine nucleoside phosphorylase). In the first case (Supplementary Fig. 6A), FTSite was unable to find the correct ligand binding site using all 3 top ranked sites. This is perhaps unsurprising, given that only the Fv fragment is mapped without the Fc domains of the light and heavy chains. The absence of these domains could result in the misidentifications of the domain-domain interface as potential ligand binding site. As for 1GCG (glucose/galactose receptor; Supplementary Fig. 6B) and 1ULA (purine nucleoside phosphorylase; Supplementary Fig. 6C), the ligands are in tight and enclosed binding sites in the bound

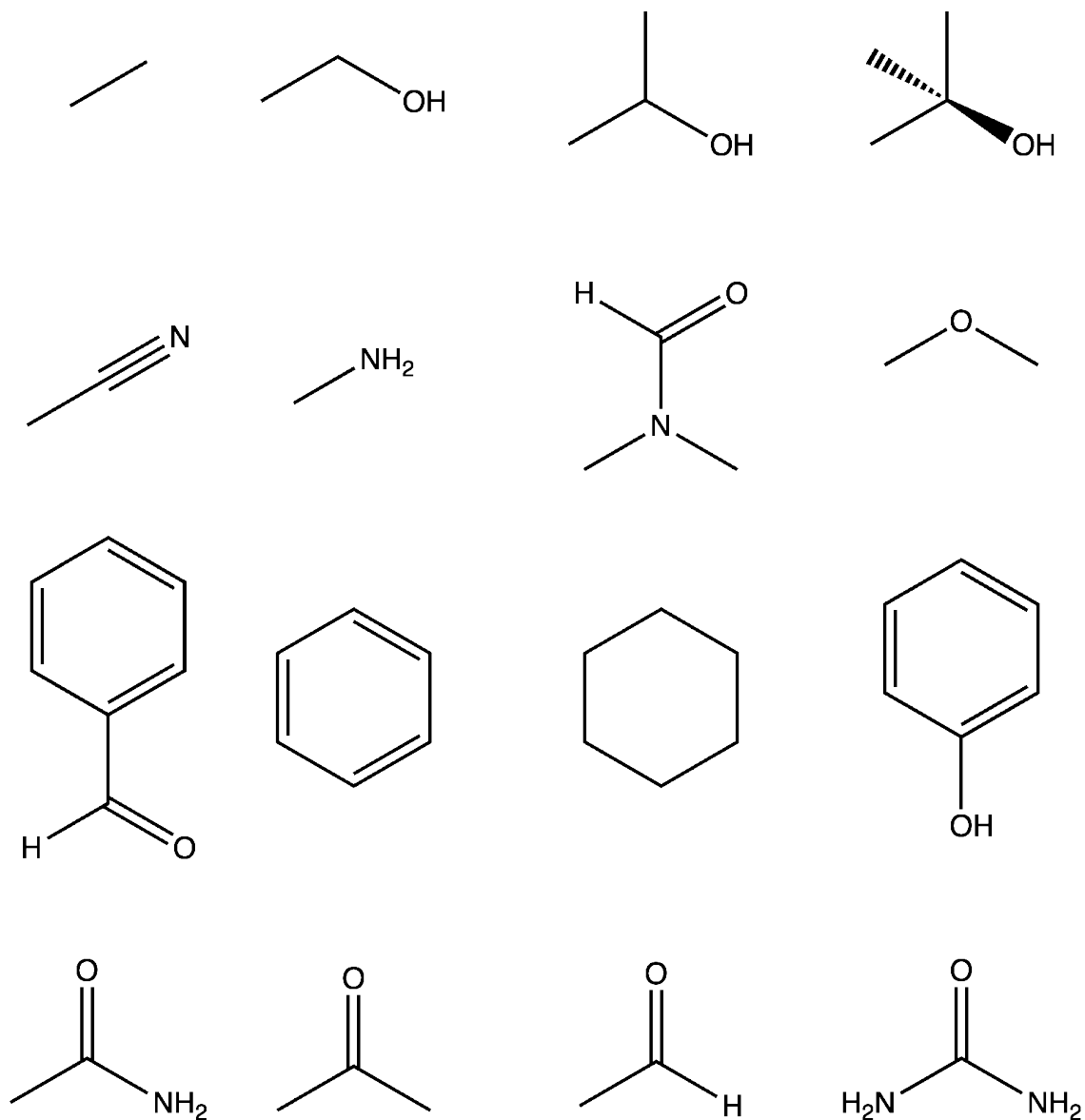
forms of the proteins. Tight and enclosed ligand binding sites present a challenge to the mapping when placing the probes because of repulsive interactions. In spite of these challenges, FTSite was able to identify the correct ligand binding site using the 2nd predicted site for both these proteins. Perhaps more interestingly, for both proteins the top ranked predictions are very close to the entrance of the ligand binding sites. The clustering of small molecular probes in these regions suggests that the amino acid residues involved are likely to be responsible for recruiting a ligand to the binding site. We have reported similar results for some haloalkane dehalogenases that have a small active site deeply buried in the protein, and an intermediate binding site on the surface^{26,27}.

The FTSite server

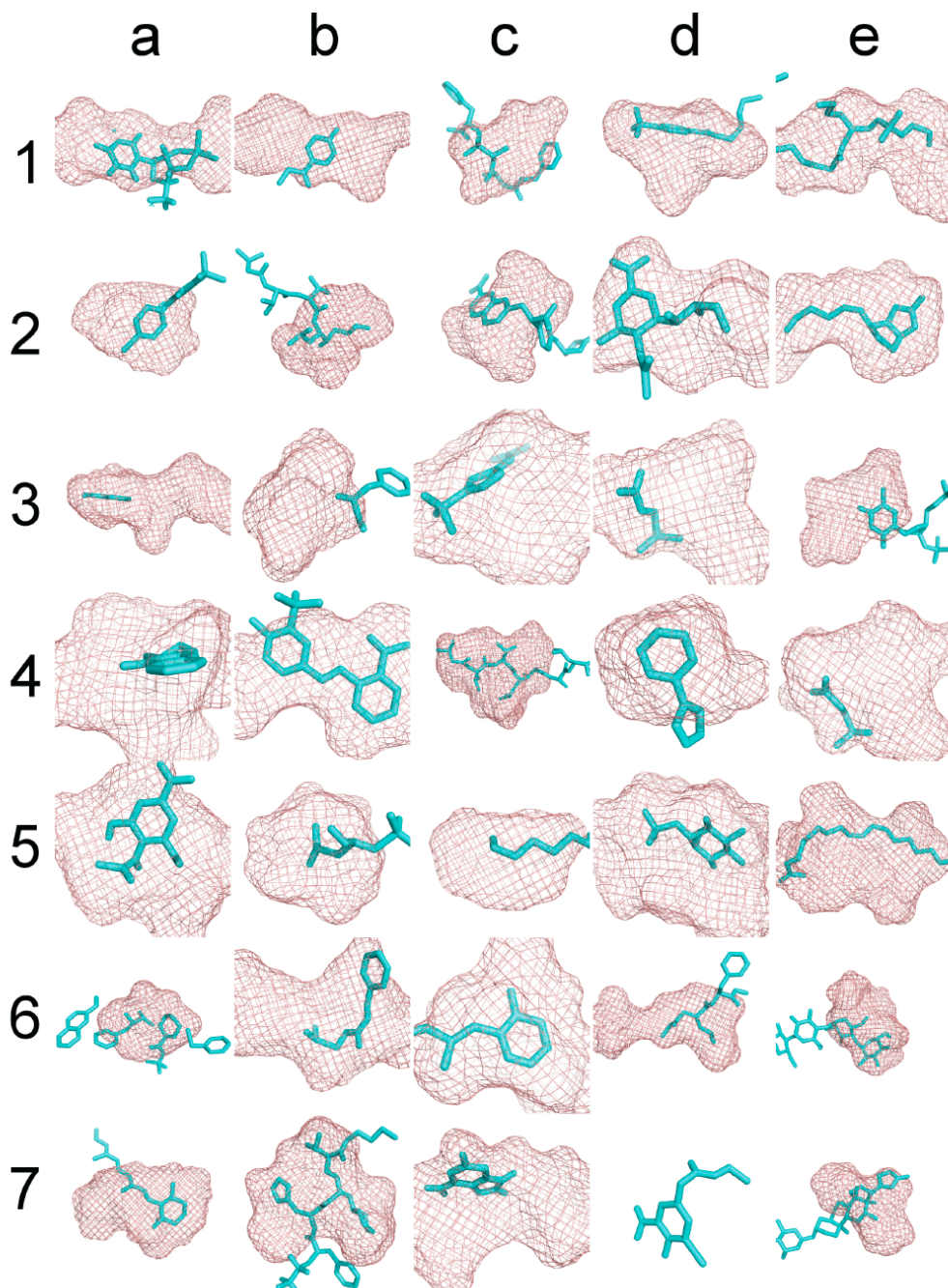
FTSite is available as a server at <http://ftsites.bu.edu>. Supplementary Figure 7A shows the interface for submitting a protein to the server. The user has the option of submitting a protein from the Protein Data Bank by specifying the PDB ID and the chain ID, or uploading a PDB file to the server. The server automatically discards all HETATM records from the PDB file, thereby removing all ligands including water molecules prior to the calculations. The user has the option of retaining metal ions, by prepending the letter “h” to the ion name, and specifying this as an additional chain during submission. For instance, to include a zinc ion, the user would specify an additional PDB chain “hzn” after the Chain ID of the protein. As shown in Supplementary Figure 7A, the user has decided to submit for analysis chain A of the tyrosine protein kinase crystal structure (PDB ID 3lck). Upon successful submission, the user should see “Success” as shown in Supplementary Figure 7B.

After the analysis is completed, the server sends an email to the user with a PyMol session and also a link to a web interface for viewing the results. The PyMol session contains the submitted protein structure with the ligand binding sites, and residues contacting each ligand binding site. As shown in Supplementary Figure 8, the web interface uses a Java Applet which is a modified version of OpenAstexViewer²⁸. The mesh representation of the sites and the sticks representation of the residues contacting the sites can be turned on and off along with other graphical representations of the protein. Right-clicking on the viewer opens a menu for saving images. The residues contacting each ligand binding site are also listed.

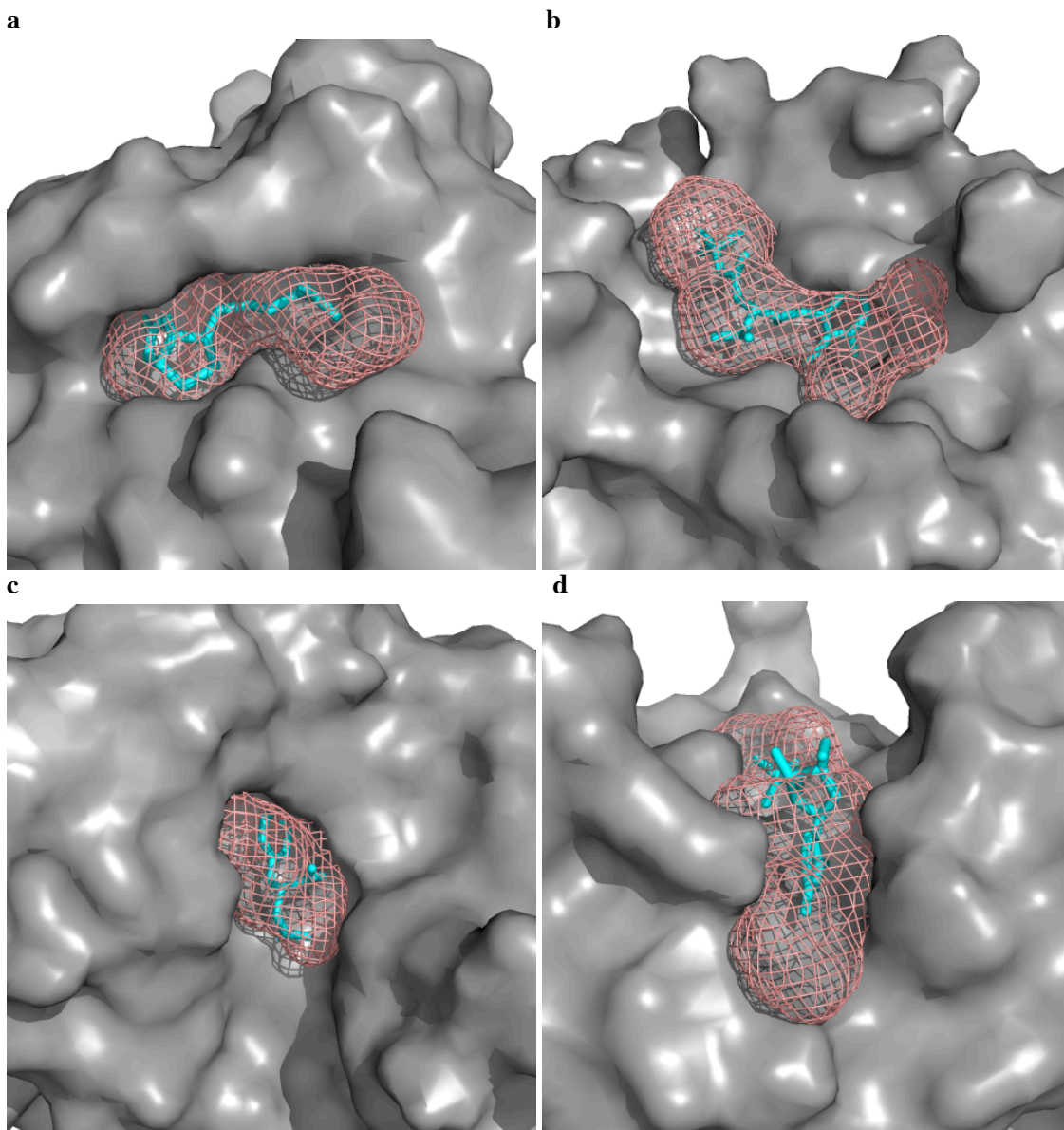
Supplementary Figure 1. Small molecules used as probes by FTSite to identify ligand binding sites. The probes are selected to include a range of chemical moieties such as hydrophobic molecules coupled with polar functional groups and aromatic compounds. Prior studies have shown this selection to be capable of providing good characterization of ligand binding sub-sites, which contribute to a disproportionate amount of binding free energy on the protein surface.



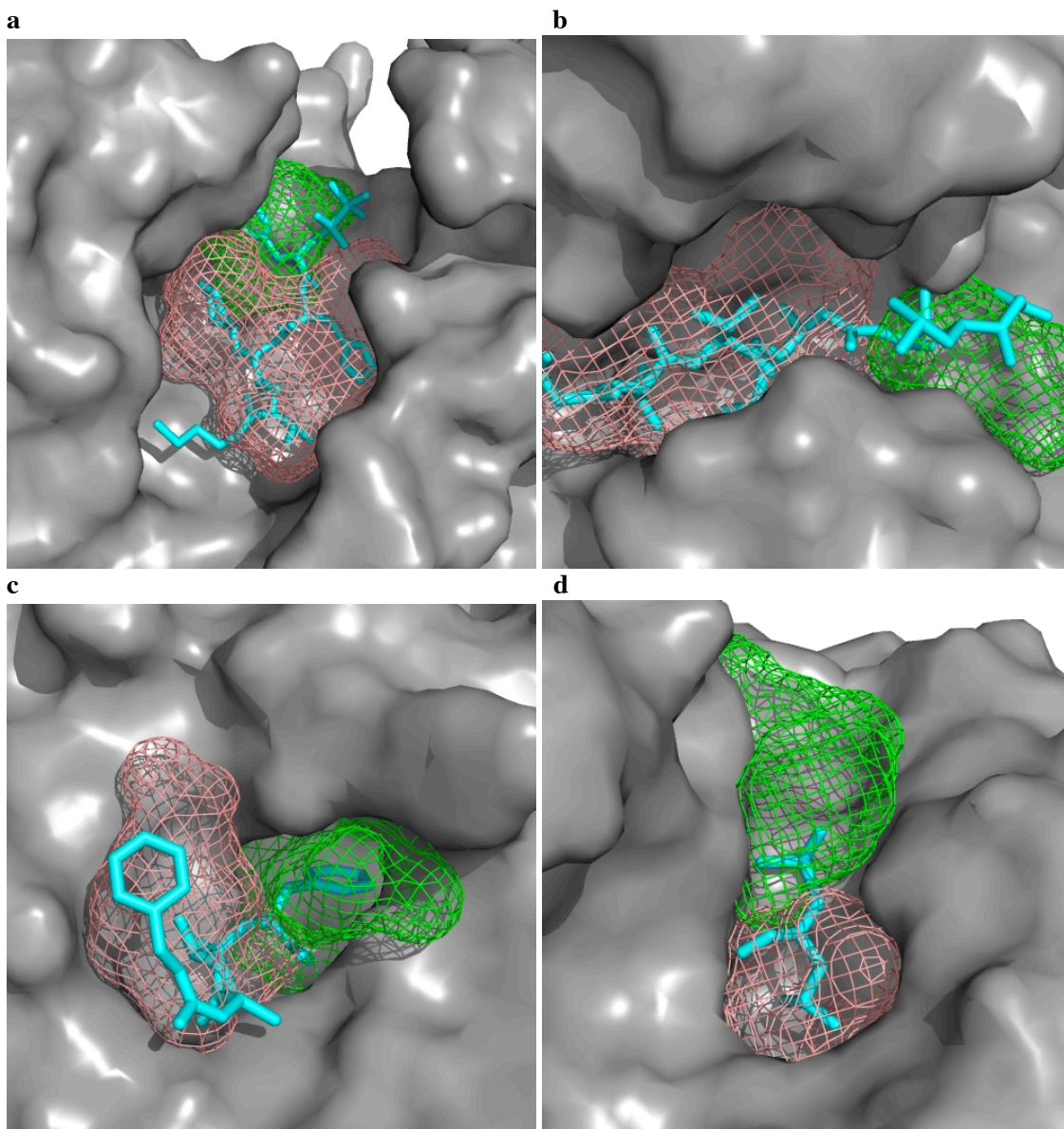
Supplementary Figure 2. Top ranked predictions of binding sites (shown in mesh representation), identified by FTSite in the unbound structures of the QSiteFinder test set. Ligands from the bound structures are superimposed for reference, and are shown in sticks representation. The corresponding unbound and bound structures for each panel can be referenced using Supplementary Table 2 (e.g. ‘1a’ refers to unbound structure PDB ID: 7RAT, and the bound structure PDB ID: 6RSA).



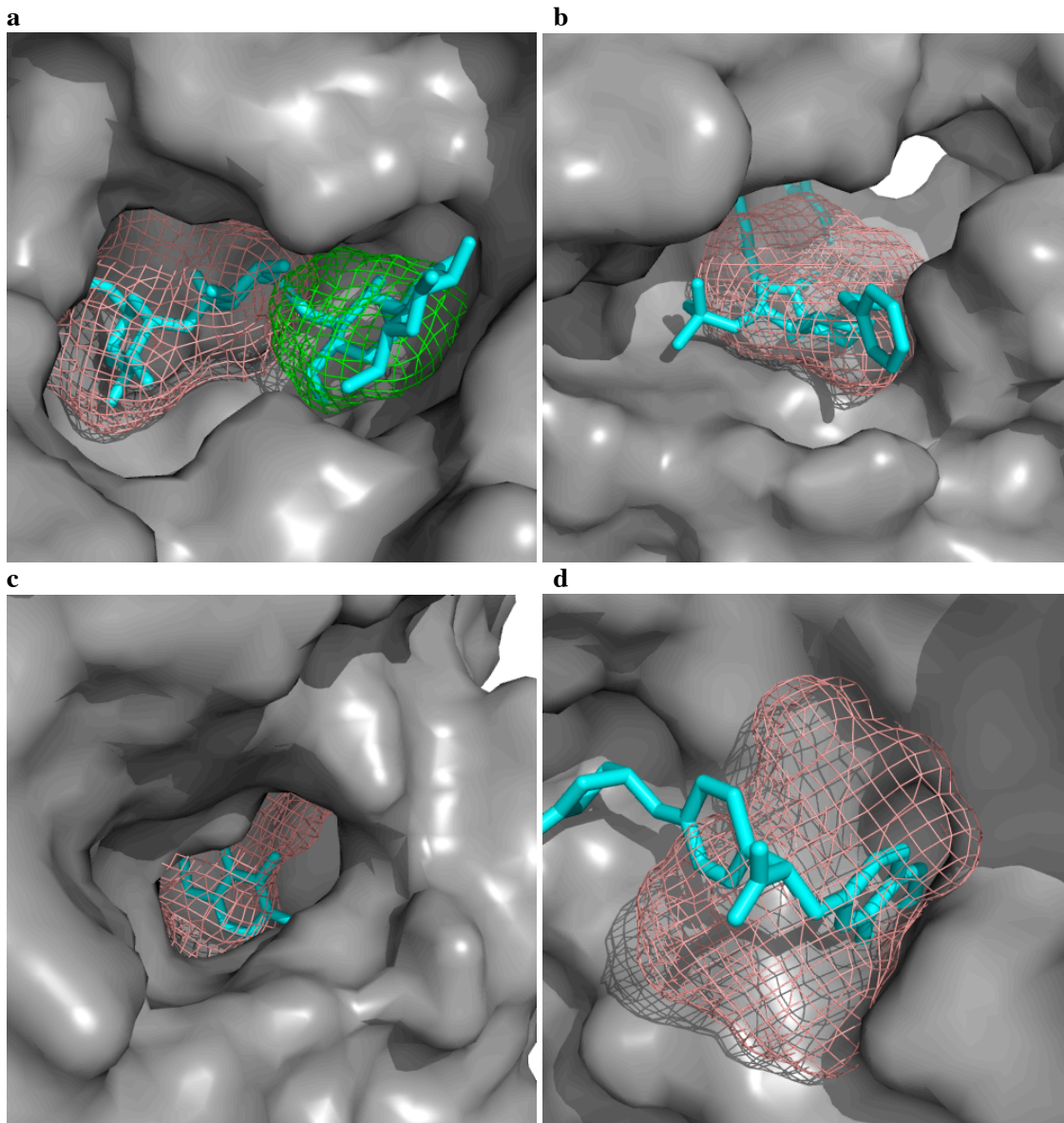
Supplementary Figure 3. The top ranked FTSite prediction of the ligand binding site, based on the unbound protein structure, has excellent coverage of the ligand (shown in sticks representation) from the bound form of the protein. The examples are listed in the following format: protein name followed by (in parenthesis) the PDB ID of the unbound structure and the PDB ID of the bound protein structure. **(a)** Streptavidin (2RTA and 1STP); **(b)** elastase (1ESA and 1INC); **(c)** protease (1NPC and 1HYT); and **(d)** Ribonuclease A (8RAT and 1ROB).



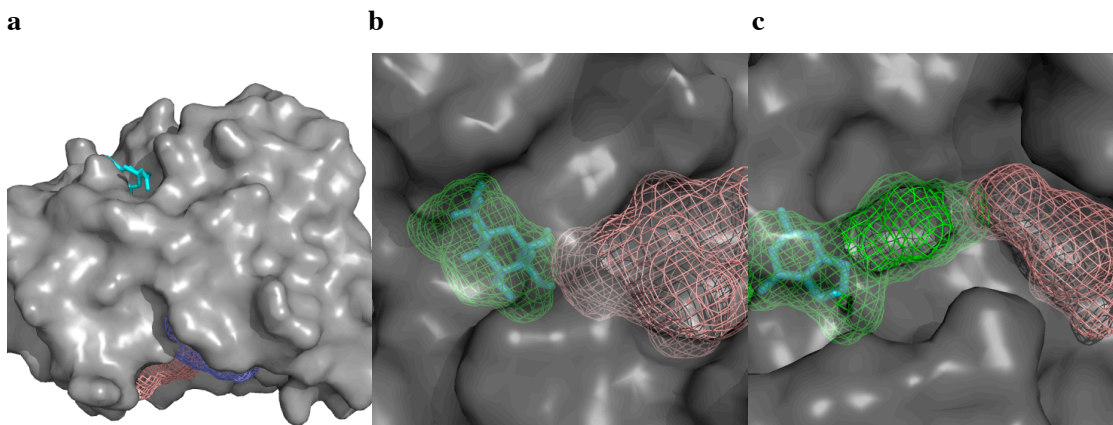
Supplementary Figure 4. Cases in which the ligands are large and/or are branched. In the following examples FTSite required the two top ranked predictions of the binding site to cover the entire ligand (shown in sticks representation). The Rank 1 prediction is colored in salmon, and the Rank 2 in green. As in the previous figure, the examples are listed in the following format: protein name followed by (in parenthesis) the PDB ID of the unbound structure and the PDB ID of the bound protein structure. **(a)** Renin (1BBS and 1RNE); **(b)** pepsin (1PSN and 1PSO) (c) carboxypeptidase (5CPA and 6CPA); and **(d)** thermolysin (1L3F and 2TMN).



Supplementary Figure 5. Cases in which FTSite succeeded in finding the ligand binding sites using the top ranked ligand binding site but other methods had challenges. These cases typically demonstrate conformational changes in which the ligand binding site is well formed only in the ligand-bound structure of the protein. The ligands are shown in sticks representation. The first putative ligand-binding site is colored in salmon, and the second is in green. As in the previous figures, the examples are listed in the following format: protein name followed by (in parenthesis) the PDB ID of the unbound structure and the PDB ID of the bound protein structure. **(a)** Beta-amylase (1BYA and 1BYB); **(b)** HIV-2 protease (1HSI and 1IDA); **(c)** acetylcholinesterase (1QIF and 1ACJ); and **(d)** trypsinogen (2TGA and 1MTW).



Supplementary Figure 6. Cases in which FTSite was unable to identify the ligand binding site using only the top ranked prediction. The ligands are shown in sticks representation. The first predicted site is colored in salmon, second is in green, and the third is in purple. As in the previous figures, the examples are listed in the following format: protein name followed by (in parenthesis) the PDB ID of the unbound structure and the PDB ID of the bound protein structure. **(a)** Antibody Fv fragment (1A6U and 1A6W); **(b)** glucose/galactose receptor (1GCG and 1GCA); and **(c)** purine nucleoside phosphorylase (1ULA and 1ULB).



Supplementary Figure 7. User interface for the FTSite server **(a)** PDB submission interface. The user can provide a job name, specify a PDB ID or upload a PDB file, provide the chain ID, and the email address for notification when the analysis is done. **(b)** An example of a successful submission.

a

Job Name:

Tyrosine Protein Kinase

PDB ID:

3lck

--or--

Upload PDB File:

Choose File No file chosen

PDB Chains:

A

E-mail:

john@example.com

Find My Binding Site

b

Success:

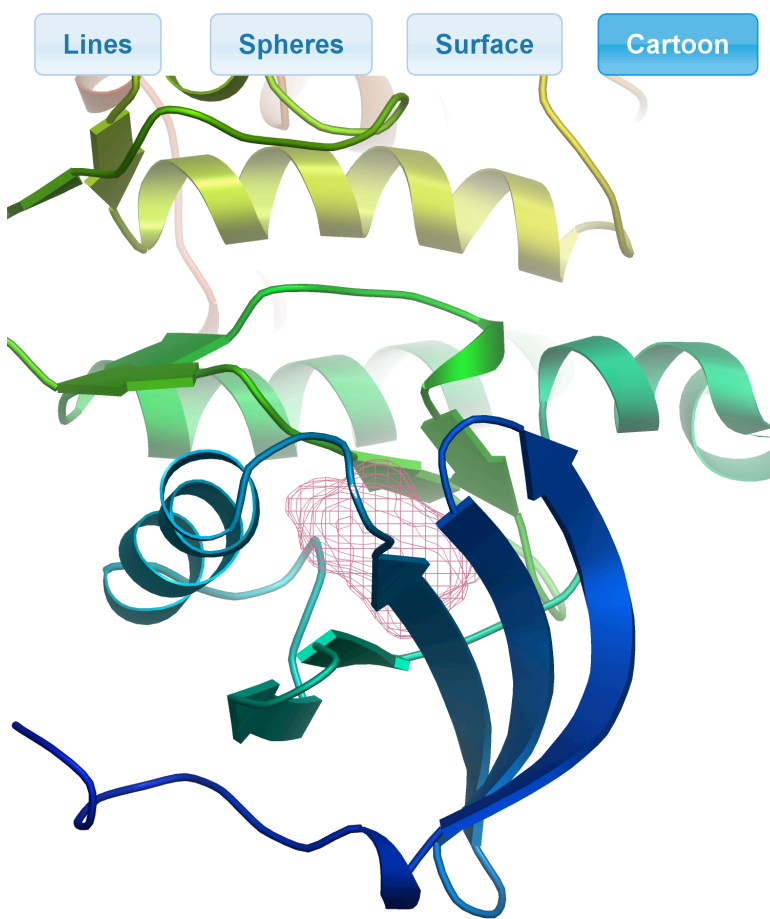
Job Submitted

Supplementary Figure 8 – Online interface for the user to view the results after the analysis is done. The interface uses a Java applet based on OpenAstexViewer.²⁴ The user has the option of generating various graphical representations of the protein, the putative ligand binding sites, and amino acid residues in contact with the respective sites. The user has the option of downloading a PyMol session as well.

3lck

[Download Pymol Session](#)

Lines
Spheres
Surface
Cartoon



Site 1

Mesh

Residues

VAL 259
TRP 260
ALA 271
VAL 272
LYS 273
GLU 288
MET 292
VAL 301
ILE 314
ILE 315
THR 316
GLU 317
LEU 371
ALA 381
ASP 382

▶ Site 2

▶ Site 3

Left Mouse Rotate

Shift + Left Scale

Ctrl + Left Translate

 +,- Clipping

 r Reset view

Supplementary Table 1 – Contact energies for atoms from the protein and atoms from the 16 small molecule probes used by FTSite

Probe Type	Atom	N	C ^α	C	O	GC ^α	C ^β	K N ^δ	K C ^δ	DO ^δ	R N ^η	N N ^η	R N ^ε	SO ^γ	H N ^ε	Y C ^δ	F C ^δ	LC ^δ	CS ^γ	
acetamide	N1	-0.04	-0.10	-0.04	-0.07	0.00	0.33	-0.06	0.10	-0.03	0.23	0.10	0.26	-0.36	-0.69	-0.23	1.13	1.26	-0.63	
	C2	-0.04	0.05	-0.03	-0.08	-0.08	0.26	-0.06	-0.07	-0.20	0.38	0.14	0.14	-0.33	-0.32	-0.18	0.76	0.62	-0.32	
	C3	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	O4	-0.07	-0.23	-0.08	-0.18	-0.10	0.18	-0.06	-0.03	0.17	-0.11	0.07	0.13	-0.01	-0.23	-0.00	0.36	0.57	0.05	
	C1	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56	
	C2	0.23	0.21	0.38	-0.11	0.15	0.14	-0.08	0.05	-0.08	-0.20	0.03	0.22	-0.30	-1.76	-0.23	0.95	1.56	-1.92	
	N	0.23	0.21	0.38	-0.11	0.15	0.14	-0.08	0.05	-0.08	-0.20	0.03	0.22	-0.30	-1.76	-0.23	0.95	1.56	-1.92	
acetone	C	-0.04	0.05	-0.03	-0.08	-0.08	0.26	-0.06	-0.07	-0.20	0.38	0.14	0.14	-0.33	-0.32	-0.18	0.76	0.62	-0.32	
	O	0.10	0.17	0.14	0.07	0.04	-0.16	0.17	-0.24	-0.16	0.03	-0.12	-0.17	0.18	0.24	0.17	-0.66	-0.81	-0.03	
	CH1	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	C1	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56	
	C2	-0.04	0.05	-0.03	-0.08	-0.08	0.26	-0.06	-0.07	-0.20	0.38	0.14	0.14	-0.33	-0.32	-0.18	0.76	0.62	-0.32	
	O1	-0.07	-0.23	-0.08	-0.18	-0.10	0.18	-0.06	-0.03	0.17	-0.11	0.07	0.13	-0.01	-0.23	-0.00	0.36	0.57	0.05	
	N1	-0.36	-0.34	-0.33	-0.01	-0.19	0.24	-0.05	0.04	0.40	-0.30	0.18	-0.05	0.32	0.52	0.10	-0.70	-0.82	0.66	
benzaldehyde	N1	-0.36	-0.34	-0.33	-0.01	-0.19	0.24	-0.05	0.04	0.40	-0.30	0.18	-0.05	0.32	0.52	0.10	-0.70	-0.82	0.66	
	C1	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	C2	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	C3	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	C4	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	C5	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	C6	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
benzohydroxyacetone	C7	-0.04	0.05	-0.03	-0.08	-0.08	0.26	-0.06	-0.07	-0.20	0.38	0.14	0.14	-0.33	-0.32	-0.18	0.76	0.62	-0.32	
	O1	-0.07	-0.23	-0.08	-0.18	-0.10	0.18	-0.06	-0.03	0.17	-0.11	0.07	0.13	-0.01	-0.23	-0.00	0.36	0.57	0.05	
	CG	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	CD1	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	CE1	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	CZ	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	CE2	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
benzene	CD2	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	C	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	isobutanol	C1	1.26	1.42	0.62	0.57	0.75	-2.32	0.63	-0.25	-1.99	1.56	-0.81	-0.94	-0.82	1.32	-0.45	-1.20	-3.49	-0.09
	isobutanol	C2	1.26	1.42	0.62	0.57	0.75	-2.32	0.63	-0.25	-1.99	1.56	-0.81	-0.94	-0.82	1.32	-0.45	-1.20	-3.49	-0.09
	isobutanol	C3	1.26	1.42	0.62	0.57	0.75	-2.32	0.63	-0.25	-1.99	1.56	-0.81	-0.94	-0.82	1.32	-0.45	-1.20	-3.49	-0.09
	cyclohexane	C1	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
	cyclohexane	C2	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
NN-dimethylformamide	cyclohexane	C3	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
	cyclohexane	C6	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
	cyclohexane	N3	0.10	0.17	0.14	0.07	0.04	-0.16	0.17	-0.24	-0.16	0.03	-0.12	-0.17	0.18	0.24	0.17	-0.66	-0.81	-0.03
	cyclohexane	O4	0.10	0.17	0.14	0.07	0.04	-0.16	0.17	-0.24	-0.16	0.03	-0.12	-0.17	0.18	0.24	0.17	-0.66	-0.81	-0.03
	cyclohexane	C3	1.26	1.42	0.62	0.57	0.75	-2.32	0.63	-0.25	-1.99	1.56	-0.81	-0.94	-0.82	1.32	-0.45	-1.20	-3.49	-0.09
	cyclohexane	C1	1.26	1.42	0.62	0.57	0.75	-2.32	0.63	-0.25	-1.99	1.56	-0.81	-0.94	-0.82	1.32	-0.45	-1.20	-3.49	-0.09
	cyclohexane	C1	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
NN-dimethylformamide	cyclohexane	O1	-0.07	-0.23	-0.08	-0.18	-0.10	0.18	-0.06	-0.03	0.17	-0.11	0.07	0.13	-0.01	-0.23	-0.00	0.36	0.57	0.05
	cyclohexane	C1	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
	cyclohexane	C2	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
	cyclohexane	C2	-0.36	-0.34	-0.33	-0.01	-0.19	0.24	-0.05	0.04	0.40	-0.30	0.18	-0.05	0.32	0.52	0.10	-0.70	-0.82	0.66
	cyclohexane	C1	-0.10	-0.45	0.05	-0.23	-0.07	0.48	0.04	-0.01	0.03	0.21	0.17	0.33	-0.75	-0.22	1.28	1.42	-0.48	-0.48
	cyclohexane	O	-0.36	-0.34	-0.33	-0.01	-0.19	0.24	-0.05	0.04	0.40	-0.30	0.18	-0.05	0.32	0.52	0.10	-0.70	-0.82	0.66
	cyclohexane	C1	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
ethanol	C2	-0.36	-0.34	-0.33	-0.01	-0.19	0.24	-0.05	0.04	0.40	-0.30	0.18	-0.05	0.32	0.52	0.10	-0.70	-0.82	0.66	
	ethanol	C1	-0.10	-0.45	0.05	-0.23	-0.07	0.48	0.04	-0.01	0.03	0.21	0.17	0.33	-0.75	-0.22	1.28	1.42	-0.48	-0.48
	ethanol	O	-0.36	-0.34	-0.33	-0.01	-0.19	0.24	-0.05	0.04	0.40	-0.30	0.18	-0.05	0.32	0.52	0.10	-0.70	-0.82	0.66
	ethanol	C1	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
	ethanol	C2	0.33	0.48	0.26	0.18	0.09	-1.09	0.08	0.17	-0.23	0.14	-0.16	-0.47	0.24	1.15	0.04	-1.72	-2.32	0.56
	ethanol	C1	-0.23	-0.22	-0.18	-0.00	-0.16	0.04	-0.07	0.16	0.29	-0.23	0.17	-0.05	0.10	0.80	-0.12	-0.25	-0.45	0.95
	ethanol	C2	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55
phenol	C3	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55	
	phenol	C4	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55
	phenol	C5	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55
	phenol	C6	1.13	1.28	0.76	0.36	0.62	-1.72	0.36	-0.11	-1.45	0.95	-0.66	-0.49	-0.70	0.29	-0.25	-0.24	-1.20	-0.55
	phenol	O	-0.36	-0.34	-0.33	-0.01	-0.19	0.24	-0.05	0.04	0.40	-0.30	0.18	-0.05	0.32	0.52	0.			

Supplementary Table 2. Comparing the performances of FTSite to those of QSiteFinder⁹ and the method of Morita et al.¹² on the QSiteFinder test set using site coverage (SC) as the success criterion. The column “Rank” indicates the rank of the prediction that overlaps with the ligand from the bound form of the protein. The first column corresponds to panels in Supplementary Figure 2 in which top predictions of binding sites by FTSite are displayed.

QsiteFinder Dataset								
Suppl Fig. 2 Grid			FTSite		QsiteFinder		Morita et al.	
	Unbound	Bound	Rank	SC (%)	Rank	SC (%)	Rank	SC (%)
1a	7rat	6rsa	1	93	1	70	1	90
1b	6ins	3mth	1	56	0	0	2	85
1c	5cpa	6cpa	1	93	3	72	3	90
1d	4ca2	1okm	1	79	3	100	1	52
1e	3p2p	5p2p	1	91	1	88	1	83
2a	3lck	1qpe	1	92	3	55	1	62
2b	3app	1apu	1	98	4	6	0	0
2c	2tga	1mtw	1	96	2	44	6	100
2d	2sil	2sim	1	98	2	59	3	52
2e	2rta	1stp	1	97	1	89	1	100
3a	2ptn	3ptb	1	59	1	89	1	96
3b	2ctb	2ctc	1	30	1	57	4	92
3c	2cba	2h4n	1	65	1	90	1	73
3d	1ypi	2ypi	1	60	2	35	1	47
3e	1stn	1snc	1	70	1	19	1	33
4a	1qif	1acj	1	61	1	21	1	65
4b	1pts	1srf	1	95	1	83	1	96
4c	1psn	1pso	1	94	1	74	1	97
4d	1phc	1phd	1	93	1	78	1	87
4e	1pdy	1pdz	1	51	3	66	1	49
5a	1nna	1ivd	1	93	2	87	1	88
5b	1l3f	2tmn	1	97	1	67	1	87
5c	1krn	2pk4	1	100	1	85	1	100
5d	1ime	1imb	1	76	1	57	1	40
5e	1lfb	1icn	1	71	1	54	1	57
6a	1hsi	1ida	1	100	7	6	0	0
6b	1djb	1blh	1	79	2	73	1	47
6c	1chg	3gch	1	91	2	78	2	1
6d	1cge	1hfc	1	74	1	77	1	89
6e	1bya	1byb	1	85	4	56	0	0
7a	1brq	1rbp	1	86	1	61	1	61
7b	1bbs	1rne	1	99	1	88	1	94
7c	1ahc	1mrg	1	59	1	41	1	57
7d	1a6u	1a6w	0	0	1	94	1	99
7e	1a4j	1igj	1	92	1	49	1	45

Supplementary Table 3. Prediction of ligand binding site for the proteins in the LIGSITE^{CSC} test set. Column 3 shows the rank of the predicted ligand binding site overlapping with the ligand from the bound form of the protein. The distances from the center of geometry of the predicted ligand binding site to the closest ligand atom are shown in Column 4. Site coverage (SC) and ligand coverage (LC) values are also listed. Note that in majority of the cases FTSite provides high-quality predictions, overlapping large fractions of the ligand.

LIGSITE ^{CSC} set					
				Coverage	
Unbound	Bound	Rank	Distance	SC (%)	LC (%)
7rat	6rsa	1	1.2	93	90
6ins	3mth	1	0.7	56	100
5cpa	7cpa	1	0.9	99	63
4ca2	1okm	1	1.4	79	89
3p2p	5p2p	1	1.7	91	89
3lck	1qpe	1	1.2	92	83
3app	1apu	1	0.5	98	58
2tga	1mtw	1	1.3	96	73
2sil	2sim	1	1.2	98	100
2ctb	2ctc	1	3.5	30	50
2cba	2h4n	1	1.3	65	100
1ypi	2ypi	1	2.9	60	100
1stn	1snc	1	2.2	70	44
1qif	1acj	1	1.2	61	93
1pts	1srf	1	0.4	95	100
1psn	1pso	1	2.0	94	69
1phc	1phd	1	0.6	93	100
1pdy	1pdz	1	2.7	51	100
1nna	1ivd	1	0.9	93	100
1l3f	2tmn	1	0.5	97	83
1krn	2pk4	1	0.6	100	100
1ime	1imb	1	1.3	76	100
1ifb	2ifb	1	2.1	69	100
1hsi	1ida	1	1.1	100	75
1djb	1blh	1	1.3	79	100
1chg	3gch	1	1.4	91	100
1cge	1hfc	1	1.0	74	76
1bya	1byb	1	1.0	85	53
1brq	1rbp	1	0.7	86	81
1bbs	1rne	1	0.8	99	80
1ahc	1mrg	1	1.7	59	100
1a6u	1a6w	-	-	-	-
1a4j	1igj	1	0.6	92	65
8rat	1rob	1	1.1	93	95
8adh	1cd0	1	2.1	65	41
5dfr	4dfr	1	1.8	67	64

3tms	1bid	1	3.9	25	30
3ptn	3ptb	1	1.8	61	100
3phv	4phv	1	0.7	100	28
2fbp	1fbp	1	0.7	99	88
2ctv	5cna	1	0.7	100	100
1ula	1ulb	2	2.7	48	100
1swb	1stp	1	0.8	93	100
1npc	1hyt	1	1.3	90	100
1hxf	1dwd	1	0.8	99	84
1hel	1hew	1	0.7	80	72
1gcg	1gca	2	0.7	100	100
1esa	1inc	1	1.0	91	100

Supplementary References

1. Berman, H.M., *et al.* (2000) The Protein Data Bank, *Nucleic Acids Res*, **28**, 235-242.
2. Brooks, B.R., *et al.* (1983) CHARMM: A program for macromolecular energy, minimization, and dynamics calculations, *J Comp Chem*, **4**, 187-217.
3. Schaefer, M. and Karplus, M. (1996) A Comprehensive Analytical Treatment of Continuum Electrostatics, *J Phys Chem*, **100**, 1578–1599.
4. Chuang, G.Y., *et al.* (2008) DARS (Decoys As the Reference State) potentials for protein-protein docking, *Biophys J*, **95**, 4217-4227.
5. Kozakov, D., *et al.* (2006) PIPER: an FFT-based protein docking program with pairwise potentials, *Proteins*, **65**, 392-406.
6. Lindemann, S.R., Yershova, A. and LaValle, S.M. (2004) Incremental grid sampling strategies in robotics. In *Proceedings of the Sixth International Workshop on the Algorithmic Foundations of Robotics*. Springer, Berlin/Heidelberg, Zeist, Netherlands, 313-328.
7. Brenke, R., *et al.* (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques, *Bioinformatics*, **25**, 621-627.
8. Young, T., *et al.* (2007) Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding, *Proc Natl Acad Sci U S A*, **104**, 808-813.
9. Laurie, A.T. and Jackson, R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites, *Bioinformatics*, **21**, 1908-1916.
10. Huang, B. and Schroeder, M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation, *BMC Struct Biol*, **6**, 19.
11. Brady, G.P., Jr. and Stouten, P.F. (2000) Fast prediction and visualization of protein binding pockets with PASS, *J Comput Aided Mol Des*, **14**, 383-401.
12. Morita, M., Nakamura, S. and Shimizu, K. (2008) Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures, *Proteins*, **73**, 468-479.
13. Laskowski, R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions, *J Mol Graph*, **13**, 323-330, 307-328.
14. Levitt, D.G. and Banaszak, L.J. (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids, *J Mol Graph*, **10**, 2.
15. Hendlich, M., Rippmann, F. and Barnickel, G. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins, *J Mol Graph Model*, **15**, 359-363, 389.

16. Liang, J., Edelsbrunner, H. and Woodward, C. (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design, *Protein Sci*, **7**, 1884-1897.
17. Halperin, I., Wolfson, H. and Nussinov, R. (2003) SiteLight: Binding-site prediction using phage display libraries, *Protein Sci*, **12**, 1344-1359.
18. An, J., Totrov, M. and Abagyan, R. (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes, *Mol Cell Proteomics*, **4**, 752-761.
19. Nayal, M. and Honig, B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites, *Proteins*, **63**, 892-906.
20. Le Guilloux, V., Schmidtke, P. and Tuffery, P. (2009) Fpocket: an open source platform for ligand pocket detection, *BMC Bioinformatics*, **10**, 168.
21. Capra, J.A., *et al.* (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, *PLoS Comput Biol*, **5**, e1000585.
22. Weisel, M., Proschak, E. and Schneider, G. (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors, *Chem Cent J*, **1**, 7.
23. Volkamer, A., *et al.* (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets, *J Chem Inf Model*, **50**, 2041-2052.
24. Tripathi, A. and Kellogg, G.E. (2010) A novel and efficient tool for locating and characterizing protein cavities and binding sites, *Proteins*, **78**, 825-842.
25. Mehio, W., *et al.* (2010) Identification of protein binding surfaces using surface triplet propensities, *Bioinformatics*, **26**, 2549-2555.
26. Silberstein, M., *et al.* (2003) Identification of substrate binding sites in enzymes by computational solvent mapping, *J Mol Biol*, **332**, 1095-1113.
27. Silberstein, M., Damborsky, J. and Vajda, S. (2007) Exploring the binding sites of the haloalkane dehalogenase Dh1A from *Xanthobacter autotrophicus* GJ10, *Biochemistry*, **46**, 9239-9249.
28. Hartshorn, M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design, *J Comput Aided Mol Des*, **16**, 871-881.